

Guadalupe: a browser design for heterogeneous hardware

Zhen Wang[†], Felix Xiaozhu Lin[†], Lin Zhong[†], and Mansoor Chishtie[‡]

[†]Rice University, Houston, TX [‡]Texas Instruments, Dallas, TX

ABSTRACT

Mobile systems are embracing heterogeneous architectures by getting more types of cores and more specialized cores, which allows applications to be faster and more efficient. We aim at exploiting the hardware heterogeneity from the browser without requiring any changes to either the OS or the web applications. Our design, Guadalupe, can use hardware processing units with different degrees of capability for matched browser services. It starts with a weak hardware unit, determines if and when a strong unit is needed, and seamlessly migrates to the strong one when necessary. Guadalupe not only makes more computing resources available to mobile web browsing but also improves its energy proportionality. Based on Chrome for Android and TI OMAP4, We provide a prototype browser implementation for resource loading and rendering. Compared to Chrome for Android, we show that Guadalupe browser for rendering can increase other 3D application’s frame rate by up to 767% and save 4.7% of the entire system’s energy consumption. More importantly, by using the two cases, we demonstrate that Guadalupe creates the great opportunity for many browser services to get better resource utilization and energy proportionality by exploiting hardware heterogeneity.

1. INTRODUCTION

By making the operating system and hardware *transparent* to web application¹ developers, a web browser has evolved into a powerful platform for content and application distribution. Recent development in hardware, especially mobile system hardware, however, challenges this key advantage of web applications as versus native applications.

As integrated circuits are hitting the power wall, modern computer systems, from servers to smartphones, are embracing *heterogeneity* in their hardware by adding processing units of various degrees of specialization and processing capability. First of all, the added processing units increase the computational resources available, allowing better performance for multiprocessing

systems. Furthermore, with heterogeneity, a computer system can not only execute a task on hardware customized for it with much higher energy *efficiency* but also match the hardware capability with the task workload for improved energy *proportionality*. To exploit hardware heterogeneity, native application developers either directly use the APIs or library associated with a specialized hardware unit, e.g., [13], or provide “hints” to the underpinning operating system (OS), e.g., [21].

Requiring web applications to do the same will unfortunately break the much valued system and hardware transparency of the web. Therefore, in this work, we ask: *can web applications leverage heterogeneous hardware transparently?* Our answer is a browser design called *Guadalupe*. Guadalupe recognizes two orthogonal dimensions of hardware heterogeneity: specialization and capability. It allows browser designers to define a *mapping pod*, which is a set of browser functions that can be mapped onto a group of hardware units of similar specialization, or a *hardware specialization group*. After the static mapping, Guadalupe leverages a browser’s run-time knowledge about web applications to identify the hardware unit with the suitable capability in the hardware specialization group for the mapping pod. It starts the mapping pod on the weak unit of the group but timely migrates it to a stronger one by demand at the run time. Guadalupe provides an efficient optimization to reduce the performance and energy overhead of such migrations.

Guadalupe is a design point, or a small design region from a rather large design space for exploiting heterogeneous hardware in the browser. In this paper, we provide the design principles that help us derive Guadalupe and describe a prototype implementation of it based on the open-source Chromium browser. The prototype maps two key mapping pods, i.e., resource loading and rendering, to the two extremes of specialization, i.e., general-purpose processors and graphics accelerators, respectively. Using a tablet development system for OMAP4 mobile application processor (SoC or System-on-Chip) from Texas Instruments, we demonstrate how Guadalupe realizes the key performance and

¹In this work, we use *web application* to refer to both more traditional, static web pages and more modern, interactive and dynamic ones.

efficiency benefits of hardware heterogeneity. We show that Guadalupe browser for rendering can reduce the 3D accelerator usage by up to 75% and frees it for potential 3D tasks from other applications. On emerging mobile systems where multiple applications can run concurrently, e.g., Microsoft Surface and Samsung Galaxy Note, the resources freed by Guadalupe can increase the other 3D application’s frame rate by 18.5% to 767%. At the same time, Guadalupe browser reduces the energy consumption of the entire system by 4.7%.

With Guadalupe and its implementation, we make the following contributions:

- A set of design principles for exploiting hardware heterogeneity for web applications in a transparent manner.
- A specific browser design, Guadalupe, that follows the principles in exploiting heterogeneous hardware.
- An implementation of Guadalupe based on Chromium and TI OMAP4 mobile SoC. We experimentally show Guadalupe improves resource utilization and energy proportionality for web applications.

To the best of our knowledge, Guadalupe is the first to explore the mapping between available hardware resources, in particular heterogeneous ones, and the mapping pods. Our effort is orthogonal to related work that incorporates more OS functions, e.g., [29], and embrace parallelism, e.g., [5, 26, 27]. As these proposals bring more tasks into the browser and extract parallel tasks from browser services, they provide new mapping pods to consider for heterogeneous hardware units and increase the potential benefits of Guadalupe.

In this work, we present Guadalupe in the context of mobile systems because mobile systems are the leading platform in embracing heterogeneous hardware. We do expect its design principles will be applicable to browsers on more powerful systems when the latter slowly though inevitably embrace specialized hardware units of various strength.

The rest of the paper is organized as follows. Section 2 introduces the background of heterogeneous architecture and browser internals. Section 3 exemplifies the benefit of exploiting hardware heterogeneity with two case studies. Section 4 describes the principles and the design of Guadalupe. Section 5 describes the prototype implementation of Guadalupe design. Section 6 presents the evaluation of Guadalupe browser for the two cases. Section 7 discuss the related work. Section 8 concludes the paper.

2. BACKGROUND

The key objective of Guadalupe is to execute a mapping pod on the most suitable hardware unit. There-

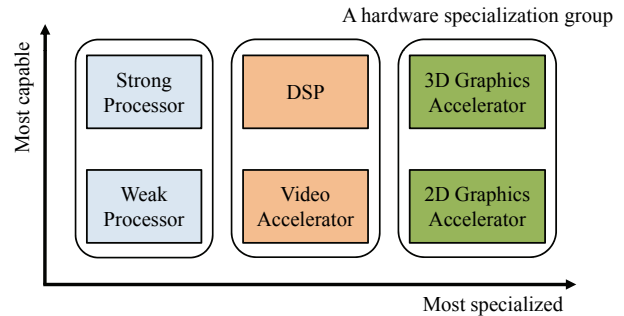


Figure 1: Various heterogeneous hardware processing units with different specializations and capabilities on mobile SoCs

fore, we first provide the background for heterogeneous hardware architectures and browser internals.

2.1 Heterogeneous Architecture

As modern computing systems are often power and energy constrained, heterogeneous architecture has become a popular strategy for higher power efficiency. Mobile systems have been a leading platform in embracing this strategy.

Heterogeneity includes two orthogonal dimensions as illustrated by Figure 1. First, a heterogeneous architecture often employs processing units of various degrees of *specialization*, from general-purpose processors like ARM cores to processors with special instruction set (ISA), e.g., digital signal processor (DSP), to application-specific accelerators treated as I/O devices by the OS, e.g., graphics accelerator, as shown along the X axis of the figure. A specialized unit is often optimized for a specialized workload and can deliver the same performance with higher efficiency than general-purpose processors by orders of magnitude [15]. For example, TI OMAP4470 mobile application processor [36] has both ARM Cortex-A9 and M-3 cores, audio back end, DSP subsystem, image and video accelerator high-definition subsystem, display subsystem, face detect module, image subsystem, and graphics accelerator.

Second, a heterogeneous architecture can employ processing units with different capabilities for the same specialization, as shown along the Y axis of Figure 1. More capable units have more functionalities and better performance, but they may also incur higher power consumption. We will use *strong* and *weak* to refer to more capable and less capable processing units of the same specialization in the rest of this paper, respectively. The dimension of capability is necessary because tasks of the same type and specialization may have a wide range of workload; and low-power and low-performance unit is necessary for light workload. For example, light-weight workload cannot fully utilize a powerful processor’s architecture features, e.g., a deep

pipeline, superscalar, speculative execution and large cache. Low-power processors can execute the light-weight workload with much higher efficiency [22]. The emerging ARM big.LITTLE architecture [14] attests to this strategy with general-purpose cores with different capabilities. TI OMAP4 also provides both Cortex-A9 and Cortex-M3 ARM cores as well as graphics accelerators of 2D and 3D capabilities.

The two dimensions of heterogeneity are dealt with differently, as will be discussed in Section 4: the browser designer statically maps the mapping pod to a hardware specialization; and Guadalupe dynamically determines the capability requirement. We will demonstrate the benefits of Guadalupe with the two extremes of the specialization dimension: general-purpose processors and graphics accelerators.

2.2 Browser Internals

A browser is both an application and an “OS”. A browser is an application running on the OS and needs many system resources such as CPU, graphics accelerators, memory, storage, network and I/O devices like touch screen. As an application, it accesses these resources through system calls provided by the underlying OS. However, a browser is also more than an application and starts to serve as a platform or “OS” for web applications, e.g., providing the interface to access hardware units [43] and enforcing the boundary between web applications [6].

However, a browser has more knowledge about its web applications than a traditional OS has about its native applications. An OS knows very limited information of a native application and the information is passed from the native application through a well-defined interface consisted of system calls. In contrast, the boundary between web applications and the browser is blurred. The browser fetches source code of a web application, parses it and creates web application state inside the browser. Therefore, the browser has almost full knowledge about web applications running on it, including their data structures and run-time behavior. This gives the browser unique opportunities to determine the best hardware to execute tasks on behalf of web applications. In contrast, the developers of a native application often have to explicitly give “hints” to the OS to determine the best hardware for execution, e.g., [2, 13, 21, 30].

A browser represents the *state* of a web application with several tree structures, namely, DOM tree, Render tree, and RenderLayer tree. The *DOM tree* stores the web application content, e.g., text and images. The *Render tree* has a one-to-one mapping to DOM tree’s visible nodes and knows how to render them. The nodes in the Render tree are divided into several groups and each group corresponds to one render layer. The *RenderLayer* tree ensures the correct rendering order among the render layers.

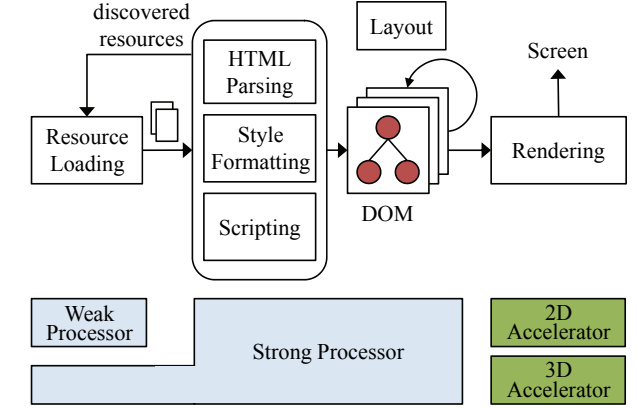


Figure 2: Basic browser services on a heterogeneous SoC

derLayer tree ensures the correct rendering order among the render layers.

A browser provides many services to a web application. Each service is a collection of browser functions with similar semantics, i.e., working on the same type of objects and producing similar outputs. A browser’s functions have been naturally organized into six basic services, as shown in Figure 2: resource loading, HTML parsing, style formatting, layout, scripting and rendering. Resource loading fetches resource files referenced by a web page, e.g., HTML, CSS, JavaScript, and image files. HTML Parsing processes the HTML file and generates the DOM tree to represent the web page’s content. New resources may be discovered by the parsing service and the browser will load them accordingly. Style formatting and layout calculate the styles and positions of the web page’s content, respectively. And they generate the Render tree and RenderLayer tree inside the browser. Scripting executes JavaScript code to provide enhanced user interaction with the web by manipulating the web application’s state, i.e., the tree structures in the browser. Finally, rendering shows the web page’s content onto the screen.

Apart from the six basic services, a browser also provides some add-on services, e.g., video decoding and image processing. With the evolution of the web and HTML standards, more and more add-on services will be added into the browser’s functionalities. We will discuss how a mapping pod is determined from those services and mapped onto a hardware specialization in Section 4.

3. CASE STUDIES

Various services performed by the browser can greatly benefit from heterogeneous hardware, in terms of resource utilization and energy efficiency. We next use two cases, resource loading and rendering, to exemplify the benefits and lay out the facts that motivate the

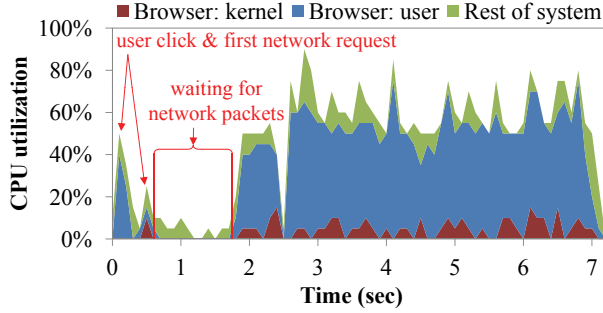


Figure 3: Measured CPU utilization when a browser is opening a CNN news page with 3G network on a smartphone. “Browser: kernel” and “Browser: user” correspond to the CPU utilization of the browser process in kernel space and user space, respectively.

Guadalupe design.

3.1 Resource Loading

Resource loading fetches a resource given its URL. When the browser opens a web page, it first requests the main resource, which is usually an HTML document. After downloading and parsing the main resource, the browser can usually discover more resources that are needed, e.g., CSS, JavaScript, and image files. They are called subresources. The browser will then fetch those subresources for additional content or page format and manipulation.

Resource loading can be parallel with other browser services because of the browser’s incremental rendering feature. Incremental rendering enables the browser to show the partially downloaded web page to the user while the browser is still loading more resources. For example, while the browser is loading subresources, the browser can layout the web page and render the partially downloaded web page onto the screen.

Resource loading is usually bounded by network latency. Mobile browsers can take 2 seconds to get the first data packet of the main resource under 3G network [44]. During this period, the browser experiences light workloads, spending most of time blocking on the network IO. Afterwards, while loading subresources, the browser starts to experience heavy workloads because subresource loading is parallel to other intensive browser services such as rendering. Figure 3 shows the CPU utilization of a browser process in opening a web page. During the early stage of page opening, i.e., before 1.8 sec, the browser incurs relatively low CPU usage, i.e., 10X lower than that of the later stage, waiting for the first a few packets. Only after that, the browser starts to consume more CPU time, in parsing resource files and rendering web contents.

TCP loopback micro-benchmark. In today’s het-

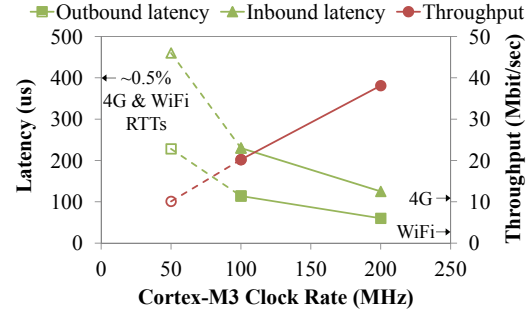


Figure 4: TCP loopback performance on Cortex-M3 of OMAP4. The solid markers are measured while the hollow ones are extrapolated. For comparison, typical wireless latency and bandwidth [17] are marked on y-axis.

erogeneous SoCs, such early-stage resource loading can be executed with typical weak processors with imperceivable performance loss. We see evidence of this by measuring the performance of TCP/IP, the heart of resource loading, on a Cortex-M3 processor of OMAP4 SoC.

In the experiment, we employ the TCP loopback benchmark: the M3 core streams 1000-byte TCP packets to and from a `loopback` interface. With stressing processors and memory, TCP loopback is a widely accepted benchmark for network stack performance. To develop the benchmark, we port `lwIP` [8], a lightweight yet full-fledged TCP/IP stack to Cortex-M3, bootstrapped by a preliminary version of our Kage kernel [22]. We disable zero-copy to include real data movement overhead. Other than that, we have not fine tuned the port due to time constraints. As the Cortex-M3 on OMAP4 only has two possible clock rates, we extrapolate the measured results in order to show the performance trend. Note that this limitation is specific to the OMAP4 platform and is not fundamental.

As shown in Figure 4, the TCP loopback benchmark implies that resource loading is able to achieve good performance on Cortex-M3. Even with M3 running at 50 MHz, one fourth of its maximum clock rate, the network stack reaches a throughput of 10 Mbps, close to the typical 13 Mbps bandwidth of today’s 4G network; with M3 running at 200 MHz, it achieves a throughput of 38.1 Mbps, which is 22% higher than the highest 4G bandwidth ever sampled by 4GTest [17]. Meanwhile, the network stack incurs at most a few hundred microseconds delay per packet. Such overhead is less than 1% of today’s wireless RTT, which is from tens to hundreds of milliseconds.

3.2 Rendering

Rendering is hardware accelerated by default [41], as

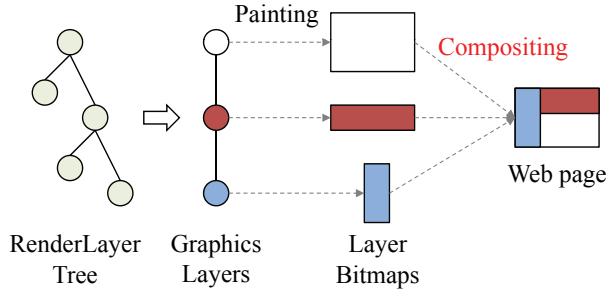


Figure 5: Hardware accelerated browser rendering. The actual hardware acceleration happens in the compositing stage.

illustrated in Figure 5. The browser first creates several graphics layers from the RenderLayer tree. Then two stages are involved: painting and compositing. In the painting stage, the browser paints each graphics layer into its own bitmap. For example, as shown Figure 5, three graphics layers are painted into three layer bitmaps. After all layers’ bitmaps are painted, in the compositing stage, the browser composites the bitmaps into one final bitmap, which is the web page.

The actual hardware acceleration happens in the compositing stage. The browser paints layer bitmaps by using CPU. Then it asks GPU to composite the bitmaps into one web page. In theory, painting stage can also be hardware accelerated, but it is very hard to map software painting commands to commands that can be understood by GPU, e.g., OpenGL commands, and the work is still on going in industry [41].

Current GPU accelerated composition only uses the 3D accelerator. But recently, the 2D accelerator is introduced to mobile SoCs [36], which can render certain web pages with much lower power consumption, while freeing the precious resource of the 3D accelerator. Therefore, browser rendering can exploit the 2D and 3D accelerators for better resource utilization and higher energy efficiency.

Our study of the Alexa top 500 web sites’ homepages [1] shows that most of them can be rendered by the 2D accelerator. In the study, we examine their rendering requirements by looking for the keywords listed in Table 1, which correspond to functions that are only provided by the 3D accelerator, but not by the 2D accelerator. As a result, out of all the 500 homepages, 449 (89.8%) can be rendered by solely using the 2D accelerator. In the rest of the paper, we will refer to them as 2D web pages and we will use 3D web pages for the other 51 homepages.

We also study the composition latency of the 2D accelerator. Figure 6 shows the cumulative distribution function (CDF) of the composition latency of the 2D accelerator for the Alexa top 500 web sites’ homepages.

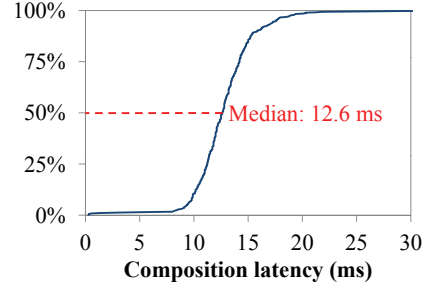


Figure 6: The CDF of the composition latency of the 2D accelerator when opening the Alexa top 500 sites’ homepages

Table 1: Common HTML tags, CSS properties and JavaScript APIs relying on functions that are only provided by the 3D accelerator

Categories	Keywords
HTML Tag	Canvas Video Object Embed
CSS Property	Animation Transform Perspective
JavaScript API	WebGL

It takes the 2D accelerator 12.6 ms to composite a web page in median, which results in a frame rate of 30 frames per second (fps), because the browser cannot finish all the rendering tasks within one display refresh interval, i.e. 16.7 ms for 60 Hz display refresh rate. However, the end user should still have the same smooth browsing experience because browsing 2D web pages does not require a very high frame rate and 30 fps is good enough, e.g., most motion pictures are filmed at 24 fps or 30 fps [45]. As for web pages that need a high frame rate, e.g., web gaming, they usually also contain CSS transformation, Canvas, and WebGL for animation. In such case, the browser have switched to use the 3D accelerator to fulfill their 3D rendering requirements.

4. GUADALUPE DESIGN

Exploiting hardware heterogeneity for web applications has a large design space. The design can be in any of the three layers: the web application, the browser or the OS. And they can choose any heterogeneous hardware processing unit freely. We identify four design principles that narrow down the design space, with decreasing granularities:

1. Make heterogeneity transparent to web developers.
2. Let the browser manage heterogeneous hardware.
3. Determine the mapping pod statically.
4. Choose hardware capability at run time.

After discussing the principles and how we derive our design, Guadalupe, based on those principles, we describe the prototype browser implementation of the design in Section 5.

4.1 Make heterogeneity transparent to web developers

We seek to free web applications from the management of heterogeneous hardware, in terms of both policy, e.g. decide which hardware to use, and mechanism, e.g. switching among hardware during execution. Our top rationale is to ease web application development: in face of today’s fast-evolving, diverse mobile platforms, it is virtually impossible for web developers to foresee users’ platforms, let alone optimize applications for platform-specific hardware heterogeneity. This rationale is also consistent with a key goal of HTML5, namely a clean separation between web application code and lower-level, platform-specific mechanisms.

Besides, current web applications can dynamically change the web content and behavior in response to the user interactions. The policy and mechanism should be able to provide transparent dynamic utilization of the heterogeneous hardware processing units in case of web application state change and require no effort from the web developers.

4.2 Let the browser manage heterogeneous hardware

Given that hardware heterogeneity are made transparent to web applications, we further argue that the browser, rather than the OS, should directly manage the heterogeneity, as will be discussed below.

Policy. The browser should always impose the *policy*, i.e., choosing the most suitable hardware for the given mapping pod, because web application information is critical in making the policy. This information includes performance hints, e.g., application behavior and future resource demands, as well as the interpretation of application internal state, e.g., its data structures. Compared to the underlying OS, the browser 1) is much closer to web applications as they run in the same address space, thus having better insight into the web application, and 2) is equipped with web-specific knowledge. Taking resource loading as an example, with the information of resource dependency, the browser is able to infer dependencies among loading requests and thus predicts CPU utilization during loading.

Mechanism. In many cases, the browser must also implement the *mechanism*, including translating application code to heterogeneous hardware primitives, supporting switch among hardware processing units during execution, etc. As hardware are increasingly specialized for applications, such mechanisms are more likely to require deeper application knowledge, which is even

less likely available to low-level, general-purpose system software such as the OS. For example, the appearance of web applications are encoded in tree structures, which have to be translated for the intended graphics accelerator for rendering. Those tree structures are complicated and thus can hardly be pushed down to general-purpose OSes.

We see a supportive evidence of this principle: existing OSes choose not to provide unified abstractions for any hardware specialization, except for general-purpose processors; rather, existing OSes treat them as separated I/O devices behind individual driver interfaces. We believe one root reason is difficulties in taking the application knowledge into the OS.

Our direct hardware management principle is also an application of the well-known end-to-end argument [32]. In our case, higher-level web software layers have richer knowledge on exploiting hardware heterogeneity. To leverage such knowledge, we push the responsibility of heterogeneity management upwards in the software stack, so that it is close to, but not into, web applications.

4.3 Determine the mapping pod statically

To utilize hardware heterogeneity, a *mapping pod* needs to be well defined. A mapping pod is a set of browser functions that can be mapped onto a hardware specialization mentioned in Section 2.1. The boundary of a mapping pod is determined by the natural boundary imposed by the hardware, but the spectrum of the boundary choices is very wide. On one extreme, the boundary can be chosen at the process level. The browser process takes the URL and shows the web page. However, it can only be mapped to the general purpose CPUs and cannot take advantage of the specialized hardware processing units available. On the other extreme, the boundary can be chosen at the instruction level. Each instruction can be mapped to different hardware. However, the strong dependency among different instructions could lead to huge overhead from choosing and switching among hardware units.

We argue that the browser designer should define the mapping pod and map it to the appropriate hardware specialization statically. Browser designers understand the browser functionalities and hardware primitives very well. During the design time, they can find the best mapping of a mapping pod and a hardware specialization to maximize performance and power efficiency. In contrast, it is too hard for the browser to figure out the correct mapping automatically.

There are two mapping pods among the six basic browser services shown in Figure 2. Resource loading is a mapping pod that can be mapped to the asymmetric processors. Rendering is a mapping pod that can be mapped to the graphics accelerators. The rest four browser services cannot benefit from any of the current

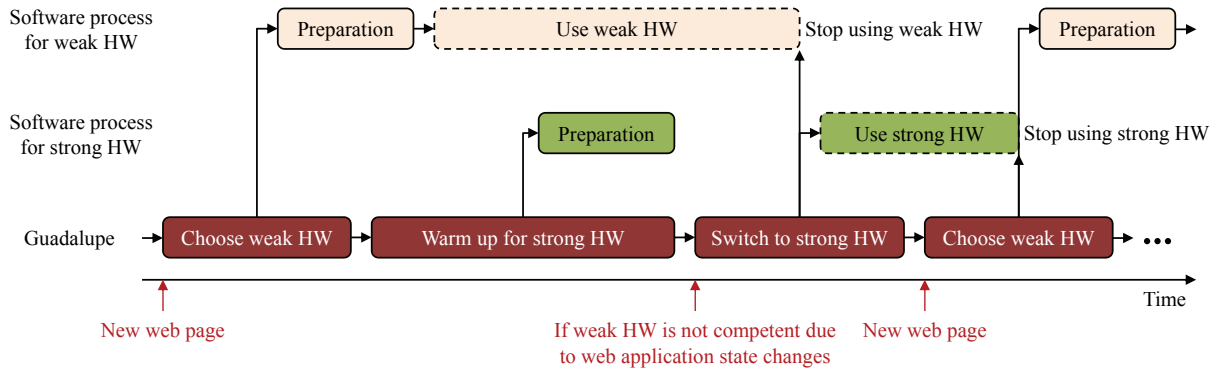


Figure 7: Guadalupe design for each mapping pod. All three software processes are running on the CPU, except the function blocks with dashed border, which run on the corresponding strong or weak hardware.

available heterogeneous hardware processing units, yet. So they belong to no mapping pod. The two add-on services, video decoding and image processing are also mapping pods and they can be mapped to the DSPs. In case a new hardware specialization emerges and can benefit other browser functionalities, a new mapping pod can be created by the browser designer to utilize those heterogeneous hardware units.

4.4 Choose hardware capability at run time

After determining the mapping pod and map it to a hardware specialization statically, the browser should choose hardware capability for the mapping pod at run time. As discussed in Section 2.1, a hardware specialization includes multiple processing units with different capabilities. The strong one has more functionalities or better performance, but consumes more energy. Based on the run-time web application state, the browser can make the best choice of the hardware capability for better performance and power efficiency.

We next use the two cases studied in Section 3 to exemplify the principle. Based on the typical workload pattern a browser has, the browser can load the main resource with the weak processor for power efficiency and load the subresources with the strong processor for good performance. As for rendering, the browser can render 2D and 3D web pages with the 2D and 3D accelerators, respectively. In case any 3D rendering requirement is added to a 2D web page, e.g., by user interaction or animation, the browser can detect the change of the application state and switch to use the 3D accelerator on demand. In this case, the browser not only makes the 3D accelerator available for other applications, but also improves energy efficiency.

4.5 Applying the principles

Guided by the four principles discussed above, we have designed Guadalupe to utilize the hardware candidate with desired capability based on the run-time dy-

namics of the web application state for better resource utilization and energy proportionality.

Guadalupe always starts from the weak hardware for each web page, and switches to the strong one on demand. The rationale is that we exploit hardware heterogeneity in the browser to make more computing resources available and improve energy proportionality. Choosing the weak hardware from the beginning for each web page is more energy efficient and frees the strong hardware for other services. In case that the weak hardware cannot provide the desired performance or cannot fulfill certain functionalities due to its limited capability, the browser will switch to use the strong hardware. Once switching to use the strong hardware, Guadalupe will not switch back to the weak one until a new web page is open, because the browser service’s requirement for the current web page will not be reduced.

Figure 7 illustrates the Guadalupe design for each mapping pod. When starting to open a web page, Guadalupe chooses the weak hardware for the mapping pod. Guadalupe monitors the web application state and checks whether the weak hardware is competent. In case the weak hardware cannot provide the desired performance or functionality, Guadalupe switches to use the strong hardware on demand and all the data structures needed by the strong hardware will be prepared. When a new web page is open, Guadalupe switches back to use the weak hardware and previously used data structures are cleared.

The key challenge in Guadalupe design is efficient switch. While getting better resource utilization and energy proportionality, Guadalupe should also switch from the weak hardware to the strong one with low overhead.

The switching overhead mainly comes from the data structure preparation for the strong hardware. We optimize Guadalupe by redundantly preparing data structures for the strong hardware, before the switch hap-

pens. So when the switch takes place, it incurs low overhead and provides a smooth transition between the two hardware units, which may not even be noticed by the user.

It is tempting to redundantly prepare the data structures for the strong hardware right before the switch is needed. However, each web application has different state, and user interaction with the web also changes the web application state dynamically, making it impossible for the browser to predict exactly when the switch is going to happen. Therefore, we design Guadalupe to prepares the required data structures for the weak and strong hardware processing units simultaneously. Those redundantly prepared data structures may not be used by the strong hardware. But in case the switching is needed, the data structures needed by the strong hardware is guaranteed to be ready, leading to low switching overhead.

Redundant preparation ensures the transparent and smooth switching. But it also incurs several overheads. The browser needs more CPU power and memory to prepare and store the redundant data structures. For the resource loading case, the redundantly prepared data structures are the URLs, which are simple and small. For the rendering case, Guadalupe needs to prepare the layer bitmaps for the 3D accelerator. We evaluate redundant preparation in Section 6 and show that the switch is fast and the overhead is small.

5. IMPLEMENTATION

We next discuss a prototype implementation of Guadalupe design: Guadalupe browser. We use TI Blaze Tablet [35] with OMAP4470 [36] as the mobile device. OMAP4470 features two types of asymmetric processors: dual Cortex-A9 and dual Cortex-M3 processors. It also has a 3D accelerator based on PowerVR SGX544 core from Imagination Technologies and a 2D accelerator based on GC320 2D core from Vivante Corporation. One can use OpenGL API to use the 3D accelerator and use BLTsville API [39] to use the 2D accelerator.

The Guadalupe browser implementation is based on Chrome for Android beta [10], which runs in Android ICS on the Blaze tablet. Chrome for Android is not fully open sourced yet [11], especially the Java side code. We are able to pull a snapshot of the Chrome for Android beta source code [7]. Combined with some other Chromium source code, we manage to modify and compile its C++ side source code and produce the shared library `libchromeview.so`. Then we push the shared library into the tablet to turn Chrome for Android beta to Guadalupe browser.

We first give an overview of the system architecture of Guadalupe browser. Then we discuss the implementation details of Guadalupe browser for resource loading and rendering.

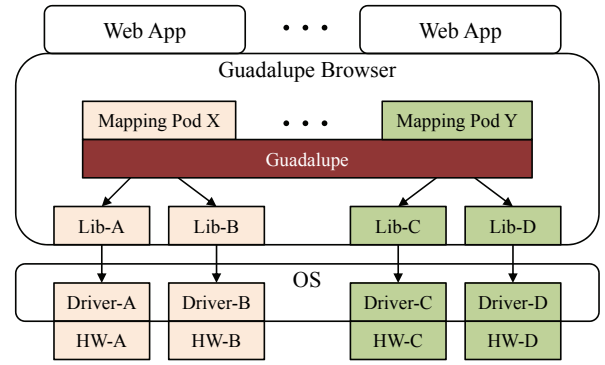


Figure 8: Guadalupe system architecture. The hardware can be either CPU or I/O devices such as graphics accelerators. Correspondingly, the driver is OS service for the CPU or device driver for I/O devices.

5.1 Overview of system architecture

Figure 8 illustrates the system architecture of the Guadalupe browser. By adding a Guadalupe layer between the mapping pod and the hardware interfaces, Guadalupe browser enables mapping pod X and Y to utilize different hardware processing units within their mapped hardware specialization. Guadalupe layer monitors the web application state, chooses the suitable hardware for the mapping pod and switches among the hardware candidates on demand. If multiple web applications are running on the browser, Guadalupe layer manages the hardware for them separately.

The hardware candidates can be CPU or specialized processing units like graphics accelerators, treated as I/O devices by the system. The hardware interface between the browser and CPU is the OS itself. The OS may also provide further abstraction to utilize heterogeneous multi-processors in the future, e.g., asymmetric processor detection and heterogeneous architecture aware system calls. The hardware interface between the browser and an I/O device is the device driver. Guadalupe browser dynamically loads the corresponding hardware library into the browser’s address space and uses it as the interface to talk to the device driver in the OS. If library loading is failed, Guadalupe browser will assume that the corresponding hardware is not available on the device.

5.2 Resource loading

Guadalupe browser exploits asymmetric processors for resource loading by loading the initial a few resources with the weak processor and later switch to the strong processor for subsequent resources.

In a legacy browser, resource loading invokes various network services, e.g., setting up TCP connections, transmitting and receiving packets, etc. After getting

the URLs of the resources, the browser sends all the resource requests to a HTTP library, which is implemented by Chrome for Android beta. The HTTP library keeps track of all the pending resource requests and in turn invokes the transport layer services provided by the OS through system calls.

In Guadalupe browser, the HTTP library is modified for our resource loading. When Guadalupe browser sends the URL request to the HTTP library, the resource information is embedded in the request to indicate which processor the request should be made with. In turn, our modified HTTP library creates and uses TCP connections on the weak or strong processor for loading the resource.

Guadalupe browser provides the policy to select processors for resource loading; it relies on a heterogeneity-aware OS to provide the mechanism that creates and maintains TCP connections on asymmetric processors. Such an OS, while missing as of now, is a key goal of our on-going efforts [22].

5.3 Rendering

Guadalupe browser for rendering exploits the 2D and 3D accelerators on the Blaze Tablet. We first give a brief background of rendering on Android system. Then we discuss the implementation details of Guadalupe browser for rendering.

On Android, apart from browser rendering for web pages, the browser also needs to render the application itself, i.e., the address bar, back and forward buttons, etc. And browser application rendering and web page rendering are separated. Figure 9 illustrates how Android rendering works with the browser. The center of the rendering system is SurfaceFlinger, which manages the buffers for the window system. The browser interacts with SurfaceFlinger through surface. First, the browser connects its surfaces to SurfaceFlinger. Then for each frame, the browser requests application buffers from SurfaceFlinger through the connected surfaces. After rendering the web page and the application, the browser posts the buffers to SurfaceFlinger. SurfaceFlinger takes the buffers from different applications, composites them into one frame buffer and posts it onto the screen.

When the browser is started, Guadalupe loads both BLTville library [39] and OpenGL library [20] into its address space, and places separate switching hooks for the 2D and 3D accelerators. A switching hook is basically a callback, which is invoked for later potential switch on demand. When Guadalupe browser opens a web page, it allocates the web page’s surface to the 2D accelerator for rendering. During the page opening, Guadalupe layer prepares the layer bitmaps for both 2D and 3D accelerators. Guadalupe monitors the changes of the web page state. Once 3D rendering require-

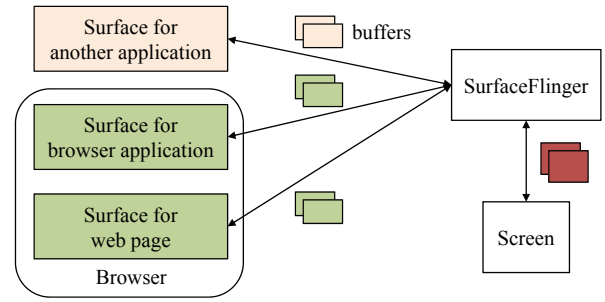


Figure 9: The Android rendering framework. Browser application rendering and web page rendering are separated.

ment is detected, Guadalupe browser switches to use the 3D accelerator by disconnecting the web page surface’s current connection to SurfaceFlinger, reallocating the surface to the 3D accelerator, and connecting the reallocated surface to SurfaceFlinger again. Note that Guadalupe manages the 2D and 3D accelerators for web page composition only, which is separated from browser application rendering.

6. EVALUATION

We evaluate the two key aspects of Guadalupe browser, rendering and resource loading. To evaluate rendering, we run Guadalupe browser on an OMAP4-based TI Blaze Tablet [35] with Android ICS. Our results show that compared to legacy browsers, Guadalupe browser provides comparable performance and better resource utilization, while incurring lower power consumption and little overhead. For evaluating resource loading, we employ a combination of estimation and micro-benchmarks on OMAP4 to show the significant benefits from Guadalupe, despite we do not yet have a complete resource loading implementation.

6.1 Rendering

Guadalupe browser for rendering is evaluated in four aspects: performance, resource utilization, efficiency and overhead. To compare with Guadalupe browser, we use Chrome for Android beta [10] and we will use mobile Chrome to refer to it in the rest of the paper. We instrument the browsers to measure the page load time, the composition latency of the 2D accelerator and overhead. We use PVRTune [18] to monitor the 3D accelerator activities and OMAPCONF [37] for bandwidth consumption of the accelerators. During the experiments, the tablet is connected to the local Ethernet network through WiFi interface. We have set up a local web page replay server [12] to remove the network variations. We first record the Alexa top 500 web sites’ homepages [1] with the web page replay server. Then we configure the tablet to use the web page replay server

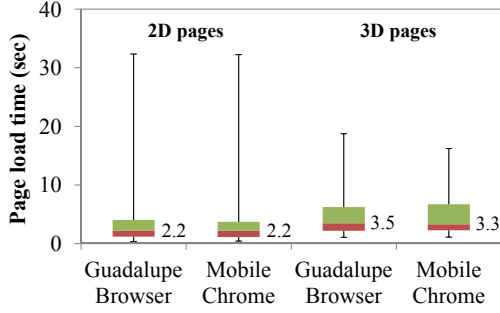


Figure 10: The web page load time of the Alexa top 500 web sites’ homepages, with data points at min, 25th percentile, median, 75th percentile and max values. The median values are labeled.

as the proxy, which always serves the resource requests from the tablet with pre-recorded resource files. We have run though the Alexa top 500 web sites’ homepages for 5 rounds for the evaluations.

While providing the same performance as mobile Chrome, Guadalupe browser makes more hardware resources available to other applications, is more power efficient, and incurs little overhead.

6.1.1 Performance

Guadalupe browser performs as good as mobile Chrome in terms of web page load time, as shown in Figure 10. For 3D web pages, Guadalupe browser starts with the 2D accelerator and switches to use the 3D accelerator after detecting 3D rendering requirements from the web application state.

With 30 fps frame rate, Guadalupe browser also provides the same smooth browsing experience, because 2D web pages does not require a very high frame rate, as discussed in Section 3.2.

6.1.2 Resource utilization

Guadalupe browser makes more hardware resources available to other applications. Figure 11 shows an example of the 3D accelerator activities when Guadalupe browser and mobile Chrome are actively compositing the web page. By utilizing the 2D accelerator, Guadalupe browser reduces the usage of the 3D accelerator by 75% and frees it for potential 3D tasks from other applications. One reason is that Guadalupe browser involves only one 3D accelerator activity for each frame, i.e., drawing the browser application. The web page composition is done by the 2D accelerator. In contrast, mobile Chrome involves two 3D accelerator activities for both browser application drawing and web page composition. The other reason is that Guadalupe browser has smaller frame rate, as discussed in Section 3.2. But even if the two browsers has the same frame rate, Guadalupe browser still reduces the 3D accelerator usage by at least

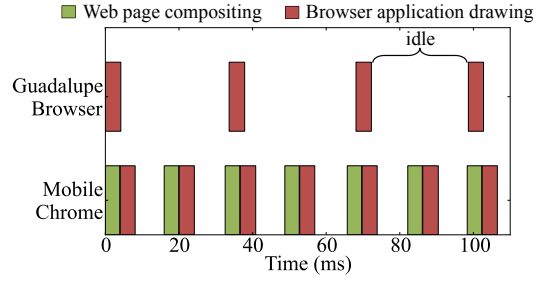


Figure 11: The 3D accelerator activities in Guadalupe browser and mobile Chrome.

50%.

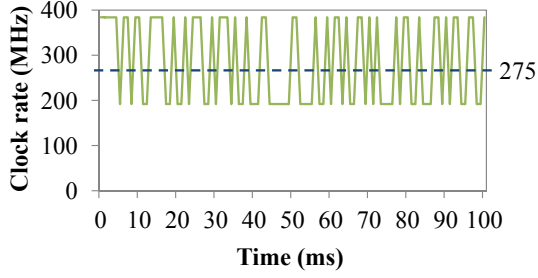
Due to the limitations of current Android system, only the front application can be actively rendered. All background applications are paused and will not be rendered, making the 3D accelerator not fully utilized. However, as future mobile devices start to support split screen and external monitor for the second front application, the idle time of the 3D accelerator freed from Guadalupe browser can be better utilized. For example, both Samsung [33] and Microsoft [28] start to sell tablets with split screen capability. While browsing a web page by using the 2D accelerator on one side of the tablet, the user can play a video, a 3D game or any 3D application on the other side of the tablet simultaneously, without too much contention for the 3D accelerator.

We estimate what frame rate the other 3D application can achieve. We use a 3D cube rotation application as the benchmark, which achieves 60 fps on Blaze Tablet. For each frame, the 3D accelerator spends 10 ms to draw the application and 6.7 ms for SurfaceFlinger composition. We assume that the browser and the benchmark can run side by side on Android and the browser’s composition need is satisfied by the 3D accelerator first. With Guadalupe browser (30 fps) running on the other side, the frame rate of the benchmark is 52 fps, which is very close to its original frame rate. However, with mobile Chrome (60 fps), the frame rate of the benchmark drops to 6 fps. Even if we set mobile Chrome’s frame rate to 30 fps, the same as Guadalupe browser, the frame rate of the benchmark still drops to 44 fps. Therefore, Guadalupe browser increases the frame rate of the other 3D application by 18% to 767%.

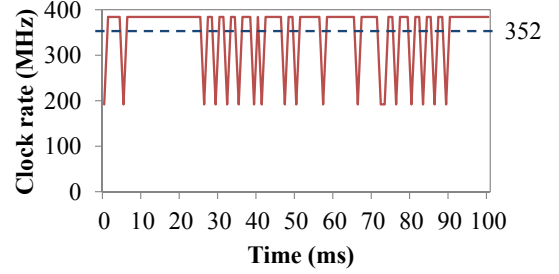
6.1.3 Efficiency

Guadalupe browser is more power efficient than mobile Chrome. As mentioned in Section 6.1.2, Guadalupe browser still needs to use the 3D accelerator to draw the application. However, it can utilize the more power efficient 2D accelerator to composite web pages.

The 2D accelerator only consumes tens of mW while

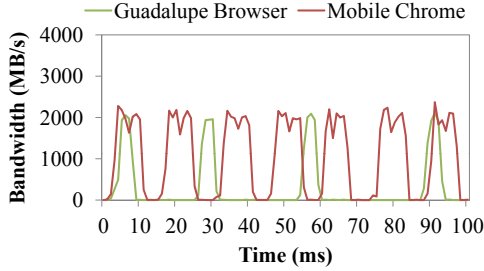


(a) Guadalupe Browser

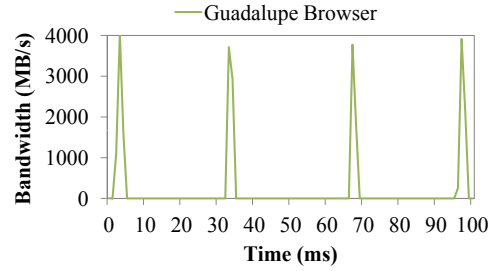


(b) Mobile Chrome

Figure 12: The clock rate of the 3D accelerator when using (a) Guadalupe browser and (b) mobile Chrome. The average clock rate is labeled in the figure.



(a) 3D accelerator



(b) 2D accelerator

Figure 13: The bandwidth consumption of (a) the 3D accelerator and (b) the 2D accelerator when using Guadalupe browser and mobile Chrome. Mobile Chrome does not use the 2D accelerator.

the 3D accelerator typically consumes several hundred mW. On OMAP4470 [36], the 3D accelerator consumes over 12 times more active power than the 2D accelerator. Due to the confidential nature of power consumption numbers of TI chip-sets involved in this study, we could not publish exact power numbers of the accelerators. However, we can compare the browsers' efficiency by showing the accelerators' clock rate, bandwidth consumption and estimated relative power consumption for composition. We also estimate how much energy Guadalupe browser can save for the entire system.

We use scrolling as the benchmark to ask the accelerators to continuously composite a 2D web page. Before each experiment, we scroll through the web page, so that all the data structures needed by the accelerators are already prepared and any further scrolling will not generate new content. Then we continuously scroll the web page for over five seconds for measurements.

The 3D accelerator in Guadalupe browser runs in a much lower average clock rate and consumes much less bandwidth, as shown in Figure 12 and Figure 13(a). Besides, the 2D accelerator in Guadalupe browser moves the web page content twice faster than the 3D accelerator in mobile Chrome, as shown in Figure 13. Furthermore, we estimate that the 2D accelerator is 6 times

more power efficient than the 3D accelerator for single frame web page composition.

We also estimate how much energy Guadalupe browser can save for the entire system. Since TI Blaze Tablet [35] is an industry prototype, its power consumption is not representative and not optimized. Instead, we measure the power consumption of Samsung Galaxy Nexus, whose SoC belongs to the same OMAP4 [38] family. Its total system power consumption is 1700 mW while browsing the web over WiFi network with full brightness of the screen. For one second of active composition, e.g., due to scrolling or 2D animation, Guadalupe browser saves 80 mJ. Therefore, Guadalupe browser saves 4.7% energy consumption of the entire mobile system. Guadalupe browser for rendering does not save much energy for the entire mobile device because other hardware components, e.g., the display, consume most of the energy. But more importantly, Guadalupe design creates the great opportunity for many other mapping pods to improve their energy proportionality.

6.1.4 Overhead

Guadalupe browser ensures efficient switch with little overhead. The switch overhead is the time between when the switch need is detected and the time when the switch is finished, which is shown in Figure 14. The

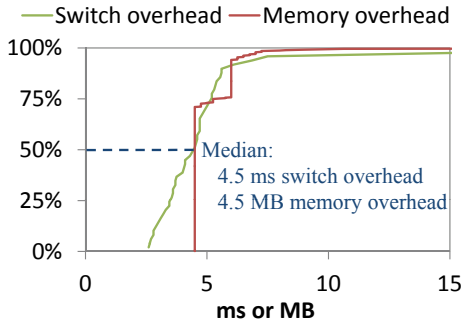


Figure 14: The CDF of the overhead of Guadalupe browser when opening the Alexa top 500 sites’ homepages.

switch of Guadalupe browser is very fast and only takes 4.5 ms in median. This is negligible comparing to the web page load time.

The overhead of redundant preparation are mainly extra memory allocation due to the data structures prepared for the hardware candidates, which is also shown in Figure 14. In median, Guadalupe browser consumes 4.5 MB for each web page, which is negligible comparing to current mobile devices’ memory size. The minimum memory overhead is also 4.5 MB because a web page needs at least one graphics layer, which consumes 4.5 MB memory. Even if the memory overhead becomes huge due to the large number of graphics layers for the web page, Guadalupe browser can always set a memory limit and stop redundant preparation to avoid excessive memory allocation.

The CPU power consumption overhead from redundant preparation is also negligible. The data structure preparation for the 2D accelerator and other browser services like parsing and layout consume over 50% CPU utilization and they have already pushes the CPU to its peak clock rate. Therefore, redundantly preparing data structures for the 3D accelerator simultaneously incurs little CPU power consumption overhead.

6.2 Resource loading

We evaluate resource loading through a combination of estimation and micro-benchmarks on TI OMAP4. We does not consider the goal of making more hardware resource available, as we expect Cortex-A9’s resource is offline to save power when Cortex-M3 performs loading.

6.2.1 Energy efficiency

As implied in Section 3, Guadalupe is able to greatly improve energy proportionality [22] by mapping resource loading to weak processors; we next estimate the energy reduction of this design, as compared to pinning resource loading on A9, i.e., the legacy case. We make two assumptions: 1) the same implementation of resource loading is mapped to either A9 or M3, and 2) in

Table 2: Estimated power and energy reduction of resource loading on Cortex-M3, as compared to on Cortex-A9

	M3			A9
Clock rate (MHz)	34.7	100	200	200
Power (mW)	1.7	7.2	19.0	22.5
Energy reduction	56.2%	39.3%	15.7%	–

considering the effect of DVFS, we use the corresponding clock rate and power scaling factors published by TI for the last generation OMAP SoC [42].

Our estimation results are shown in Table 2. In deriving the results, we first estimate the power of resource loading on A9. As resource loading is a light workload, we assume that A9 can perform it with the lowest clock rate at 200 MHz, i.e., its most energy efficient state. Given that A9 running at 1 GHz typically consumes 250 mW [3], by applying 5.8X clock scaling and 11X power scaling, we estimate that A9 running at 200 MHz consumes 22.5 mW.

We next estimate the power of resource loading on M3. Running at 200 MHz, M3 typically consumes 19 mW [4]. Applying scaling factors from [42], we estimate that M3 consumes 7.2 mW at 100 MHz and 1.7 mW at 34.7 MHz. Comparing the power consumption of two cases and taking account into the clock differences, we conclude that mapping resource loading to M3 can achieve as high as 56.2 % energy reduction as compared to pinning it on A9.

In practice, such an energy reduction will be even higher, due to processor idle periods that frequently occur in resource loading. In such short idle periods, A9 either spends high idle power (~ 11 mW [42]) or frequently enters and exits deep-sleep power state, both of which are energy hungry. In comparison, M3 has 10X less idle power (< 1 mW) while being able to perform much more efficient power state transitions, thanks to its lightweight architecture.

We stress that the above estimation must be based on mapping resource loading to both processors; it is wrong to evaluate energy efficiency of Guadalupe design by comparing the `lwip` performance on M3, as reported in Section 3, with the current Linux TCP/IP performance on A9. Unlike the heavily optimized Linux network stack, our `lwip` port is not only untuned but also has various implementation-specific limitations, e.g., maximum 64 KB TCP send buffer. Our reported `lwip` performance should only be read as an evidence showing that resource loading can be executed well on weak processors, even with such a preliminary implementation.

6.2.2 Switch Overhead

During page opening, as system resource demand ramps up, resource loading will be switched from M3 to A9;

the latter processor is expected to be in low-power state before the switch happens. The switch consists of two main steps, inter-processor interrupt and power state transition, which take 20-30 μ s and up to 2 ms, respectively [22]. As the switch happens only once in opening each web page, which typically takes \sim 2 secs in total, we believe the overhead is acceptable.

6.2.3 Data Sharing Overhead

Due to OMAP4's extreme heterogeneity for energy efficiency, no hardware cache-coherence exists between A9 and M3. Thus, in order to make sure that A9 has a consistent view of loaded resources, M3 must flush its cache before A9 can start to parse any loaded resources, an overhead that is absent in pinning resource loading on A9.

In order to estimate an upper bound of the overhead, we run a micro benchmark on M3 to periodically flush its entire 32 KB cache. Our measurement shows that the flush operation takes M3 \sim 3000 cycles, or 15 μ s, to complete. Again, as flushing happens only once in opening each web page, we think the overhead is acceptable.

7. RELATED WORK

With the advent of heterogeneous hardware architecture, OS support for heterogeneous hardware management emerges. The Linux community is working toward supporting the heterogeneous multi-processor aware scheduler [24] for ARM big.LITTLE [14] architecture. The authors of PTask [31] propose new OS abstractions to manage GPUs as shared compute resources instead of I/O devices. Renderscript [13] enables native Android applications to run general computation operations with automatic parallelization across all available processor cores, including GPU and DSP.

However, even with OS support, exploiting hardware heterogeneity still requires the knowledge of web application, thus can hardly be done by OS. Therefore, we propose that browser should manage the heterogeneous hardware directly, which is essentially an application of two generic principles: 1) the end-to-end argument [32] and 2) that the browser can be treated as the library OS [9] for web applications.

Guadalupe is the first to exploit hardware heterogeneity for web applications. It is designed to run on commodity OS, e.g., Android, on which native applications and web applications coexist. A browser OS can potentially manage hardware processing units for web applications. For example, ServiceOS [29] brings OS functions into the browser and provides secure access control and fair resource sharing mechanisms for using system resources. But it has more freedom to modify both the browser and the kernel, since web applications are the only applications in the system. Moreover, pre-

vious browser OSes, e.g., Chromium OS [40], IBOS [34] and ServiceOS [29], were not designed to utilize multiple hardware processing units because the heterogeneous architecture just starts to become pervasive on mobile SoCs in recent days.

Gibraltar [23] abstracts the interaction between web pages and hardware components with a client-server model. It uses AJAX as the hardware access protocol and its main focus is on I/O devices such as sensors. Guadalupe utilizes the existing hardware abstraction to access hardware functionality, but it is able to select the most suitable hardware candidate based on the run-time web application state.

The W3C's Device APIs Working Group [43] produces standardized APIs for web applications to access device hardware, in order to hide platform-specific hardware from web applications. Sharing a similar goal, Guadalupe provides the policy and mechanism to select the most suitable heterogeneous hardware processing unit, and thus hiding them from applications.

Application hints and profiling have been widely explored, e.g., for file buffer cache management [30] and power management [2, 16, 19, 25]. Based on hints or profiling results, the system can predict application behaviors and thus optimize for them. Generally, producing application hints requires extra development efforts and profiling requires training period before making good prediction. Fortunately, neither of them are necessary to Guadalupe, as Guadalupe is able to gain sufficient knowledge of applications by making sense of their current state.

Some researchers [5, 26, 27] have sought to parallelize the browser. They extract parallel tasks from the browser and execute them on homogeneous multi-core system. Guadalupe is orthogonal to their work and can take their parallel tasks as new mapping pods for heterogeneous resources.

8. CONCLUDING REMARKS

Guadalupe is the first endeavor to exploit the emerging hardware heterogeneity for web applications. The design utilizes the heterogeneous processing units transparently. It provides static mapping between the mapping pod and hardware specialization, and enables the browser to choose and switch among hardware processing units at run time based on web application state. We demonstrate the benefit of Guadalupe design through the prototype browser implementation for resource loading and rendering. The design not only makes more hardware resources available, but also improves energy proportionality. More importantly, Guadalupe design opens the door to all kinds of browser services, that can potentially take advantage of the heterogeneous architecture for better performance and efficiency.

9. REFERENCES

- [1] Alexa. The top 500 sites on the web.
<http://www.alexa.com/topsites>.
- [2] Manish Anand, Edmund B. Nightingale, and Jason Flinn. Ghosts in the machine: interfaces for better power management. In *Proc. USENIX/ACM Int. Conf. Mobile Systems, Applications, & Services (MobiSys)*, 2004.
- [3] ARM. Cortex-a9 processor.
<http://www.arm.com/products/processors/cortex-a/cortex-a9.php>.
- [4] ARM. An introduction to the arm cortex-m3 processor. <http://www.arm.com/files/pdf/IntroToCortex-M3.pdf>.
- [5] Carmen Badea, Mohammad R. Haghighat, Alexandru Nicolau, and Alexander V. Veidenbaum. Towards parallelizing the layout engine of firefox. In *Proc. USENIX Conf. Hot Topics in Parallelism (HotPar)*, 2010.
- [6] A. Barth, C. Jackson, C. Reis, and TGC Team. The security architecture of the chromium browser, 2008.
- [7] P. Beverloo. Bringing Google Chrome to Android.
<http://peter.sh/2012/02/bringing-google-chrome-to-android/>, 2012.
- [8] A. Dunkels. Design and implementation of the lwip tcp/ip stack. *Swedish Institute of Computer Science*, 2:77, 2001.
- [9] D.R. Engler, M.F. Kaashoek, et al. Exokernel: An operating system architecture for application-level resource management. In *ACM SIGOPS Operating Systems Review*, volume 29, pages 251–266, 1995.
- [10] Google. Chrome for Android devices.
www.google.com/chrome/android.
- [11] Google. Chrome mobile FAQ.
<https://developers.google.com/chrome/mobile/docs/faq>.
- [12] Google. Web page replay.
<http://code.google.com/p/web-page-replay/>.
- [13] Google Renderscript. <http://developer.android.com/guide/topics/renderscript>.
- [14] P. Greenhalgh. Big.LITTLE Processing with ARM Cortex-A15 & Cortex-A7. 2011.
- [15] R. Hameed, W. Qadeer, M. Wachs, O. Azizi, A. Solomatnikov, B.C. Lee, S. Richardson, C. Kozyrakis, and M. Horowitz. Understanding sources of inefficiency in general-purpose chips. *ACM SIGARCH-Computer Architecture News*, 38(3):37, 2010.
- [16] Yoshihiko Hotta, Mitsuhiro Sato, Hideaki Kimura, Satoshi Matsuoka, Taisuke Boku, and Daisuke Takahashi. Profile-based optimization of power performance by using dynamic voltage scaling on a pc cluster. In *Proc. Int. Conf. Parallel and distributed processing (IPDPS)*, 2006.
- [17] J. Huang, F. Qian, A. Gerber, Z.M. Mao, S. Sen, and O. Spatscheck. A close examination of performance and power characteristics of 4g lte networks. In *Proc. USENIX/ACM Int. Conf. Mobile Systems, Applications, & Services (MobiSys)*, pages 225–238. ACM, 2012.
- [18] Imagination Technologies. PVRTune.
<http://www.imgtec.com/powervr/insider/powervr-pvrtune.asp>.
- [19] N. Ioannou, M. Kauschke, M. Gries, and M. Cintra. Phase-based application-driven hierarchical power management on the single-chip cloud computer. In *Proc. Int. Conf. Parallel Architectures and Compilation Techniques (PACT)*, 2011.
- [20] Khronos Group. OpenGL.
<http://www.opengl.org>.
- [21] F.X. Lin, Z. Wang, R. LiKamWa, and L. Zhong. Reflex: using low-power processors in smartphones without knowing them. In *Proc. ACM Int. Conf. Architectural Support for Programming Languages & Operating Systems*, 2012.
- [22] F.X. Lin, Z. Wang, and L. Zhong. Supporting distributed execution of smartphone workloads on loosely coupled heterogeneous processors. In *Proc. Workshp. Power-Aware Computing and Systems (HotPower)*, 2012.
- [23] K. Lin, D.C.J. Mickens, L.Z.F. Zhao, and J. Qiu. Gibraltar: exposing hardware devices to web pages using AJAX. In *Proc. USENIX Conf. Web Application Development*, 2012.
- [24] LWN.net. Linux support for ARM big.LITTLE.
<http://lwn.net/Articles/481055/>.
- [25] Grigorios Magklis, Michael L. Scott, Greg Semeraro, David H. Albonesi, and Steven Dropsho. Profile-based dynamic voltage and frequency scaling for a multiple clock domain microprocessor. In *Proc. Int. Symp. Computer Architecture (ISCA)*, 2003.
- [26] H. Mai, S. Tang, S.T. King, C. Cascaval, and P. Montesinos. A case for parallelizing web pages. In *Proc. USENIX Conf. Hot Topics in Parallelism (HotPar)*, 2012.
- [27] Leo A. Meyerovich and Rastislav Bodik. Fast and parallel webpage layout. In *Proc. Int. Conf. World Wide Web (WWW)*, 2010.
- [28] Microsoft. Surface tablet.
<http://www.microsoft.com/Surface/en-US/surface-with-windows-rt>.
- [29] A. Moshchuk and H.J. Wang. Resource management for web applications in serviceos. Technical report, Microsoft Research, 2010.
- [30] R. H. Patterson, G. A. Gibson, E. Ginting,

- D. Stodolsky, and J. Zelenka. Informed prefetching and caching. In *Proc. ACM Symp. Operating Systems Principles*, 1995.
- [31] C.J. Rossbach, J. Currey, M. Silberstein, B. Ray, and E. Witchel. PTask: operating system abstractions to manage GPUs as compute devices. In *Proc. ACM Symp. Operating Systems Principles*, pages 233–248, 2011.
- [32] J.H. Saltzer, D.P. Reed, and D.D. Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems (TOCS)*, 2(4):277–288, 1984.
- [33] Samsung. Galaxy Note 10.1. http://www.samsung.com/global/microsite/galaxynote/note_10.1/benefits.html.
- [34] S. Tang, H. Mai, and S.T. King. Trust and protection in the illinois browser operating system. In *Proc. USENIX Conf. Operating systems design and implementation (OSDI)*, pages 1–8, 2010.
- [35] Texas Instruments. Blaze Tablet. http://omappedia.org/wiki/OMAP4_BlazeTablet.
- [36] Texas Instruments. OMAP4470. <http://www.ti.com/product/OMAP4470>.
- [37] Texas Instruments. OMAPCONF. <https://github.com/omapconf/omapconf>.
- [38] Texas Instruments. OMAP4 applications processor: Technical reference manual. <http://www.ti.com/product/OMAP4470>, 2010.
- [39] Texas Instruments BLTsville. <http://graphics.github.com/bltsville/>.
- [40] The Chromium Projects. Chromium OS. <http://www.chromium.org/chromium-os>.
- [41] The Chromium Projects. GPU accelerated compositing in Chrome. <http://dev.chromium.org/developers/design-documents/gpu-accelerated-compositing-in-chrome>.
- [42] TI. Power estimation tool. <http://www.ti.com/tool/powerest/>.
- [43] W3C Device APIs Working Group. <http://www.w3.org/2009/dap>.
- [44] Zhen Wang, Felix Xiaozhu Lin, Lin Zhong, and Mansoor Chishtie. How far can client-only solutions go for mobile browser speed? In *Proc. Int. Conf. World Wide Web (WWW)*, 2012.
- [45] Wikipedia. Frame rate. http://en.wikipedia.org/wiki/Frame_rate.